Exploration of Artificial Landscapes
Based on Random Graphs

Stephan Kopp, Christian Reidys and Peter Schuster
IMB–Jena, Beutenbergstr. 11, PO–BOX 100 813, D–07708 Jena

Abstract

The mapping of RNA–molecules to secondary structures is a good example to study principles of evolutionary optimization. Instead of using a prediction algorithm which takes into account thermodynamic parameters, we apply random graph theory to perform the mapping. Thereby we focus on statistical properties. The most prominent feature of preimages (neutral networks) is the existence of a threshold value for density and connectivity properties. The neutral networks percolate sequence space and allow for optimization by using point and pair mutations.

1 Introduction

Essential information on evolutionary optimization processes are inherent in the mapping from sequences to secondary structures in RNA. Based on the RNAfold Fontana et al. [1], Hogeweg and Hesper [4] and others have studied RNA landscapes in great detail. Instead of using a folding algorithm [3, 9] Reidys et al. have modeled the preimage of a fixed RNA secondary structure as a random graph [7]. This approach allows very fast simulations of dynamics on nets induced by the graph and leads to following results: above a certain threshold value the preimage of one secondary structure is a *dense* and *connected* network in sequence space, i.e. a *neutral network*. The threshold value is given by the fraction of *neutral neighbors*, i.e. the fraction of sequences, which can be achieved by one single point or base pair mutation from a reference sequence and fold into the same structure. The random graph approach allows an explicit description of the optimization process on these nets and gives insight into theoretical aspects of evolution [6].

The mathematical modeling of a single neutral network is now extended to complete sequence to structure mappings. The resulting networks are analyzed for density and for connectivity, i.e. the *sequence of components* is

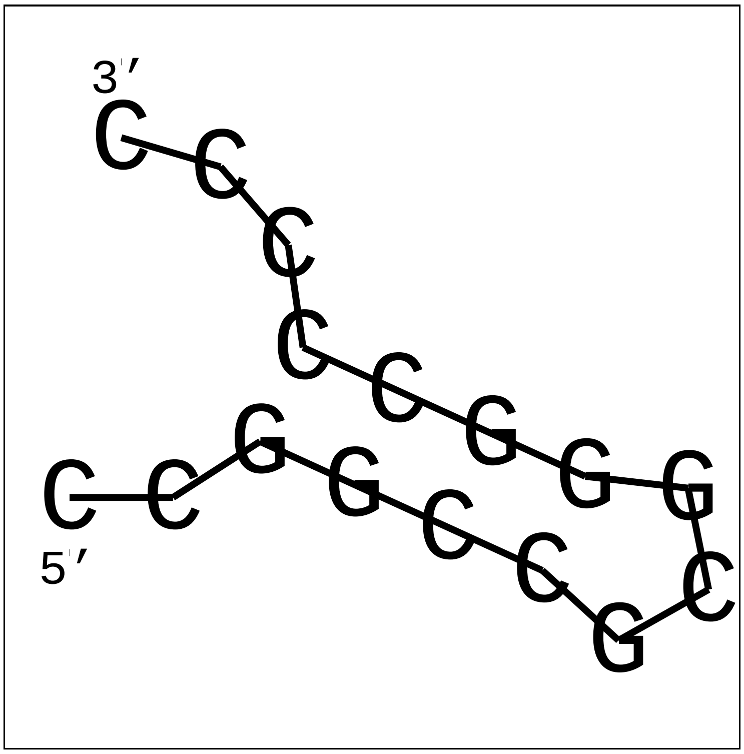

Fig. 1: Representation of a RNA secondary structure as planar graph, chain length n=16. A short notation is the bracket–dot notation [4]. A dot represents an unpaired base, an opening bracket stands for upstream paired bases, closing ones represent downstream paired bases. RNA molecules have chemically different ends. They have a natural orientation, denoted by 5' and 3'. The nucleotides are numerated, starting at the 5' end. The shown structure can be written as: 5'-..((((...))))...-3'. One compatible sequence is for instance: 5'-CCGGCCGCGGGCCCCC, which can be dissected into a subsequence corresponding to the unpaired part: 5'-CC GCG CCC and in one that corresponds to the paired part: (GC) (GC) (CG) (CG). By replacing every C by a G and vice versa, another compatible sequence can be obtained (see also text). Two structures are different, if their bracket–dot notation is different. By replacing one '(' and one ')' by a '.' one obtains another structure the given sequence is compatible with, whereas the subsequences will look different.

evaluated and the so called *shape space covering conjecture* (Schuster et al.) [8] is checked. The performance of an evolutionary search process is studied to obtain information about the number of "new" secondary structures, that are found diffusing on a neutral network [5, 6].

2 Secondary Structures of RNA Sequences

The artificial landscape is based on the biological background of RNA molecules. These *sequences* are composed of *nucleotides* or *bases* which are taken from a finite set $\mathcal{A} = \{\text{A, G, C, U}\}$. This set is called an *alphabet* with cardinality $|\mathcal{A}| =: \alpha$. In the following the sequences shall have a fixed number n of bases, i.e. a fixed *chain length*.

A natural graph structure is inherent in the *sequence space* or *hypercube*, denoted by $\mathcal{Q}_\alpha^n$. A sequence is interpreted as a vertex $\mathbf{x} = (x_1, \ldots, x_n)$ of this hypercube where $x_i \in \mathcal{A}$ for $i = 1, \ldots, n$. The Hamming distance of two sequences $d_H(\mathbf{x}, \mathbf{y})$ is defined by the number of bases which differ in $\mathbf{x}$ and $\mathbf{y}$ [2]. Two sequences $\mathbf{x}$ and $\mathbf{y}$ are said *to be adjacent* or *to be neighbors*, if $d_H(\mathbf{x}, \mathbf{y})$=1.

Due to physical and chemical properties of the bases distinct intramolecular *bases pairs* (*contacts*) are formed, e.g. (G-C), (A-U). The pairs induce a *secondary structure* (see Fig. 1) and set up the base pair alphabet $\mathcal{B} \subset \mathcal{A} \times \mathcal{A}$. To obtain the secondary structure pairs of non successive bases x_i and x_k are checked, whether they interact via hydrogen bonds or not. In the latter case they are called *unpaired*. The bond between successive bases is not considered to be a base pair. The secondary structure including the backbone is described by an adjacency matrix $A = a_{i,k}, (1 \leq, i, k \leq n)$, such that:

1. $a_{i,i+1} = 1$ for $1 \leq i < n$[1]
2. For each i there is at most one single $k \neq i \pm 1$ such that $a_{i,k} = 1$
3. If $a_{i,j} = a_{k,l} = 1$ and $i < k < j$, then $i < l < j$

The adjacency matrix A of a secondary structure s yields the *set of contacts* of s, $\Pi(s) := \{[i, k] | a_{i,k} = 1, k \neq i \pm 1\}$. Throughout the reminder of this paper we use *structure* for short, since we deal with secondary structures only. The set of all structures is called *shape space*, denoted by $\mathcal{S}_n$.

[1]The index i is also used to refer to the corresponding nucleotide x_i.

A base pairing rule for RNA molecules can be formulated as follows: If two bases x_i and x_k are in contact (i.e. $a_{i,k}$=1, and $k \neq i\pm1$), the pair (x_i, x_k) must be an element of $\mathcal{B}$. In general, any sequence $\mathbf{x}$ fulfilling the condition $(x_i, x_k) \in \mathcal{B}$ for all $[i, k] \in \Pi(s)$ is called *compatible with the structure s*. Thereby we obtain the *set of compatible sequences* (or *set of compatibles*) $C[s]$. Being compatible with a structure s is a necessary condition for a sequence to be folded into this s, but not a sufficient one.

Making use of a folding algorithm such as e.g. RNAfold [3], the secondary structures for all sequences are calculated. This algorithm takes into account thermodynamic parameters to evaluate the structure with minimal free energy. As shown by Hofacker et al. [3], the number of secondary structures Σ is much smaller than the number of sequences. Therefore many sequences must form the same structure. This observation has fundamental impact on optimization of structures by means of random walk.

3 The Artificial Landscape

Instead of using a folding algorithm, we perform the sequence to structure mapping by using Monte Carlo methods. We further restrict the alphabet to a binary one $\mathcal{A} = \{G, C\}$ inducing a pair alphabet $\mathcal{B} = \{(G, C), (C, G)\}$. Bijective mappings from the set $Ind = \{1, 2, \ldots, \Sigma\}$ into the set of structures can be made which yield $\Sigma!$ different *structure families*. Every structure family is a particular assignment of a structure to an *index*.

The artificial landscape is a mapping $f\colon \mathcal{Q}_\alpha^n \to \mathcal{S}_n$, constructed as follows:

1. A structure family is selected randomly and fixed.
2. The mapping starts with the structure corresponding to index $i = 1$.
3. Each sequence in $C[s_i]$ is mapped to the structure s_i with probability λ, if it was not mapped to a structure s_j, where $j < i$.
4. The previous step is repeated with the structure corresponding to the index $i + 1$ until $i = \Sigma$.

The sequences, which are mapped to one structure can be considered as vertices of a *random graph* (*neutral net*) $\Gamma[s]$. We use the terms *map* or *mapping* to distinguish this random map from folding routines.

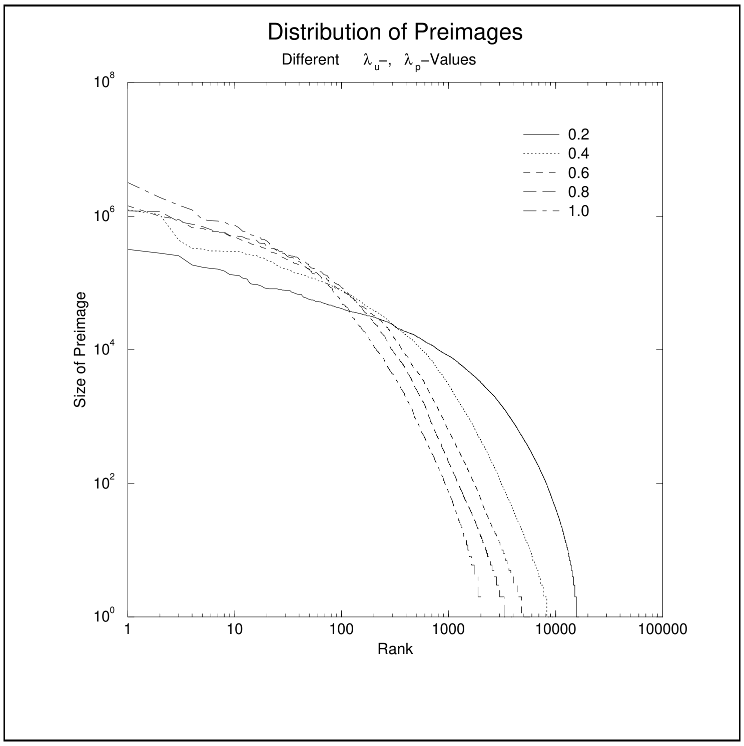

Fig. 2: Double logarithmic plot of size and rank of a graph Γ. For different λ–values the total number of structures which have a nonempty graph varies. With increasing probability λ the influence of the nonempty intersection of the sets of compatibles becomes more effective. For λ close to 1, the number of structures needed to cover $\mathcal{Q}_\alpha^n$ is less then for smaller probability values. By using the program RNAfold for all sequences in $\mathcal{Q}_2^{25}$ 18590 different structures have been discovered.

The mapping from sequence space to a structure family was performed with different probability values, using another family for each simulation. A first result of the mapping is the *rank* of the structures, i.e. the sizes of the preimages of all structures are ordered decreasingly. The ranks of the structure usually do not coincide with the indices the structures have in the family. The results are shown in Fig. 2. These plots show that the mutual influence of the $C[s]$'s depends on the probability λ but not on the structure family. There are always few structures having a large preimage and many, whose $|\Gamma|$ is small. The effect of λ on the variability in $\mathcal{S}_n$, i.e. the number

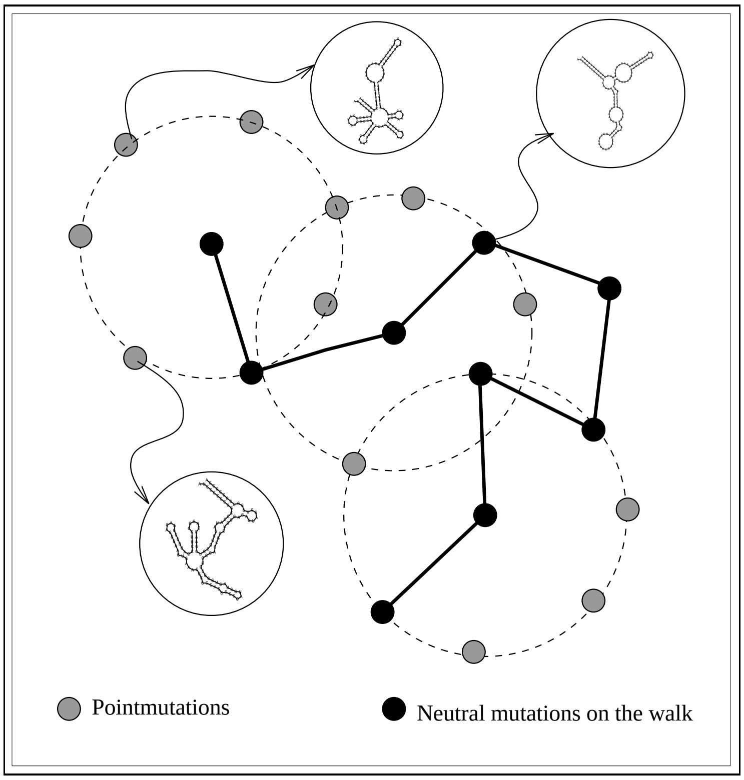

Fig. 3: A schematic presentation of a neutral walk is shown. The thick line indicates the edges of the walk, dark points represent its vertices. These vertices on the neutral walk may have Hamming distances 1 or 2. The light points on the circles always have Hamming distance 1 from the center. The vertices on this $(n-1)$–dimensional hyperplane (or ball) may be mapped to new structures. Along the neutral walk all structures will be found, whose preimage is large enough. "Rare" structures are not likely to be on any of the examined balls.

of structures with a nonempty preimage, is demonstrated.

4 Neutral Networks

The structure s consisting of n_u unpaired bases and n_p base pairs suggests to separate the compatible sequences into subsequences $\mathbf{x}_u$, coding the unpaired regions of s and $\mathbf{x}_p$, which stands for the paired part (see caption of

Fig. 1). The bases of the sequences are elements of $\mathcal{A}$ or $\mathcal{B}$ and therefore are vertices in $\mathcal{Q}_\alpha^{n_u}$ or $\mathcal{Q}_\beta^{n_p}$ respectively. The graph of all $\mathcal{C}[s]$ can be described as Cartesian product of the two hypercubes, i.e. $\mathcal{C}[s]{:=}\mathcal{Q}_\alpha^{n_u} \times \mathcal{Q}_\beta^{n_p} \subset \mathcal{Q}_\alpha^n$. (We will identify a graph with its set of vertices, since the adjacency of vertices is given in a natural way as described above.) Within the two hypercubes, the subsequences can be selected independently with probabilities λ_u and λ_p.

In general the *boundary* of a graph $\Gamma \in \mathcal{Q}$ is $\partial\Gamma := \{\mathbf{x} \in \mathcal{Q} \setminus \Gamma \mid \exists\, \mathbf{y} \in \Gamma : d_H(\mathbf{x},\mathbf{y}) = 1\}$ and the *closure* is $\overline{\Gamma} := \partial\Gamma \cup \Gamma$. A graph is called *dense* in $\mathcal{Q}$, if $\overline{\Gamma} = \mathcal{Q}$.

Rank	$\lvert\Gamma\rvert$	Structure	μ_u	μ_p	Sequence of Components
1	905355	..............(((....)))	0.249	0.234	899338, 4×4, 17×3, 213×2,
					5524×1
2	743836	((....((.......)).)).....	0.356	0.355	743759, 77×1
3	714053	....((((....)))).........	0.344	0.343	713941, 112×1
4	658860	....((...))....((.....)).	0.324	0.318	658493, 2×2, 363×1
5	575243	.....((((........))))....	0.293	0.293	574041, 3, 13×2, 1173×1
6	574192	.....((.((.....)).)).....	0.294	0.286	573073, 21×2, 1077×1
7	456412	((....((.......))...))...	0.247	0.239	452708, 22×3, 141×2, 3356×1
8	435871	.((..((........))....))..	0.243	0.232	431682, 6, 4, 17×3, 164×2,
					3800×1
9	376554	(((......((...))...)))...	0.359	0.359	376504, 50×1
10	360095	...((((.....)))).........	0.211	0.202	352793, 5, 8×4, 34×3, 355×2,
					6453×1

Tab. 1: Ranking and sequence of components for the first 10 structures in rank. The graphs are obtained from a mapping with probability parameters $\lambda_u = \lambda_p = 0.6$. The values μ show the fraction of neutral neighbors, calculated for the unpaired and paired part separately.

Assuming infinite chain length it can be shown that threshold values λ^* for the density and connectivity properties are inherent in $\Gamma_u[s]$ and $\Gamma_p[s]$ [7]. These thresholds are given by the equations $\lambda_u^* = 1 - \alpha^{1/(1-\alpha)}$. These considerations are analog for $\Gamma_p[s]$ and λ_p^* with corresponding alphabet $\mathcal{B}$. Above the threshold the subgraphs are always connected and dense in $\mathcal{Q}_\alpha^{n_u}$ and $\mathcal{Q}_\beta^{n_p}$. Below this threshold none of these properties exist.

The graph $\Gamma[s]$ is constructed by the product of the two subgraphs $\Gamma[s] = \Gamma_u[s] \times \Gamma_p[s]$. The graph Γ is dense and connected, if both subgraphs are.

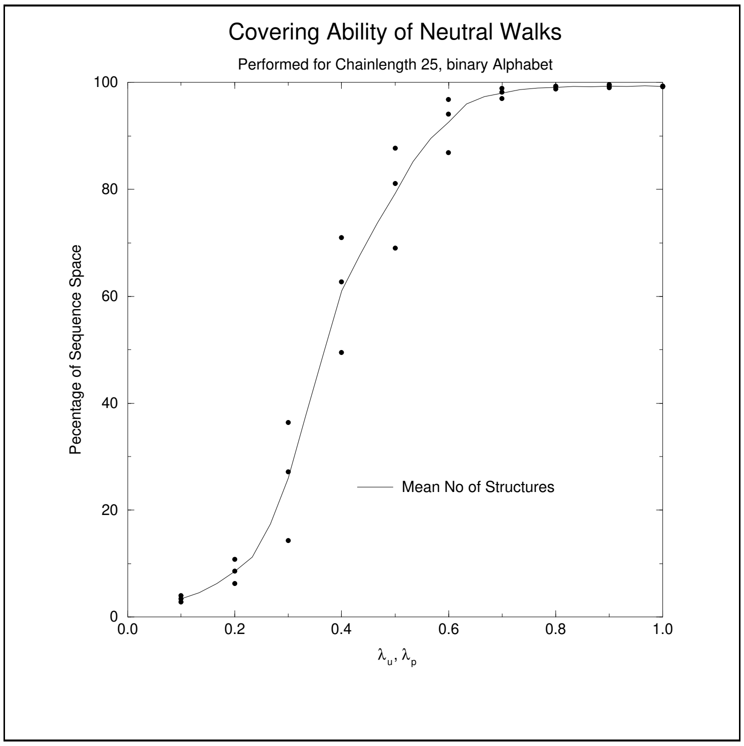

Fig. 4: 20 independent neutral walks have been performed for $\lambda_u = \lambda_p = 0.1, 0.2, \ldots, 1.0$. The mapping experiments also have been done for all λ parameters. The mean number $\mathcal{N}$ of new structures found on the walk was evaluated as well as the its standard deviation. The percentages of the sequence space covered by these structures are taken from the corresponding mapping.

For finite chain length the graphs of all structures are examined. We observe that $\Gamma[s]$ usually consist of more than one component $\chi_1 \cup \ldots \cup \chi_K = \Gamma[s]$. The most important observation is, that there exists always one component $\chi_1[s]$ with $|\chi_1[s]| > 2/3|\Gamma[s]|$. This one is called *giant component.* The sequence of components is achieved by ordering the components according to $|\chi_i[s]| > |\chi_{i+1}[s]|$. Exemplary results of graph decompositions are shown in Tab. 1.

5 Conclusions

The existance of an intersection $C[s] \cap C[s'] \neq \emptyset$ for all $s, s' \in \mathcal{S}_n$ in combination with the density property results in the fact that two neutral nets $\Gamma[s]$ and $\Gamma[s']$ come close, if the probability values are above their threshold values. This allows for optimization by means of neutral walks, where base pair mutations are considered as point mutations in $\mathcal{Q}_\beta^{n_p}$. A schematic presentation is shown in Fig. 3. While diffusing on $\Gamma[s]$ the sequences in $\partial\Gamma[s]$ are examined.

The number $\mathcal{N}$ of new structures found depends on λ_u and λ_p. The ability of these $\mathcal{N}$ structures to cover sequence space can be seen in Fig. 4. Although the variability of the structures increases with smaller λ values, the $\mathcal{N}$ most frequent structures cannot cover $\mathcal{Q}_\alpha^n$. The existence of a threshold value λ^* is demonstrated, whereas it is not exactly 0.5 due to effects of finite chain length. Indirectly, we could validate the shape space covering conjecture [8].

References

[1] W. Fontana, W. Schnabl, and P. Schuster. Physical aspects of evolutionary optimization and adaption. *Phys. Rev. A*, 40(6):3301–3321, 1989.

[2] R. Hamming. Error detecting and error correcting codes. *Bell Syst. Tech. J.*, 29:147–160, 1950.

[3] I. L. Hofacker, W. Fontana, P. F. Stadler, L. Bonhoeffer, M. Tacker, and P. Schuster. Fast Folding and Comparison of RNA Secondary Structures (the Vienna RNA Package). *Mh. Chem.*, 125(2):167–188, 1994.

[4] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucl. Acids Res.*, 12:67–74, 1984.

[5] M. Huynen, P. F. Stadler, and W. Fontana. Evolutionary dynamics of RNA and the neutral theory. *PNAS*, 1994. accepted.

[6] C. Reidys, C. V. Forst, and P. Schuster. Replication on neutral networks of RNA secondary structures. *Bull. Math. Biol.*, 1995. submitted.

[7] C. Reidys, P. F. Stadler, and P. Schuster. Generic Properties of Combinatory Maps and Neutral Networks of RNA Secondary Structures. *Bull. Math. Biol.*, 1995. submitted.

[8] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond.*, B 255:279–284, 1994.

[9] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, 9:133–148, 1981.